Вариант 2

Упражнение 1

Александрова Александра

20 02 2021

# Данные

Функция для задачи 1:

Характеристики для задачи 2:

# Задача 1.

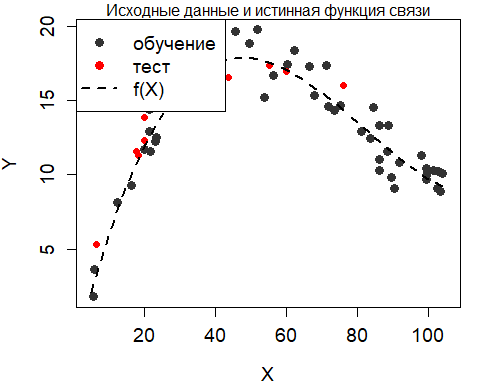
На данных своего варианта повторить три графика из первой практики, выбрав число степеней свободы как компромисс между точностью (оценкой ошибки на тестовой выборке) и простотой модели (числом степеней свободы). Все рисунки сохранить в графические файлы в формате png.

Сгенерируем X и Y линейной модели и изобразим исходные данные на графике.

#### Генерация данных

Изобразим исходные данные на графике.

###### График 1: Исходные данные на график



Изобразим исходные данные на графике.

В качестве модели используем сплайны со степенями свободы от 2 (прямая) до 40 (количество узлов равно 2/3 наблюдений). Строим модели с различным количеством степеней свободы и в каждом случае считаем среднеквадратическую ошибку модели на обучающей и тестовой выборках.

Получаем таблицу вида:

## train test   
## 1.24 1.64

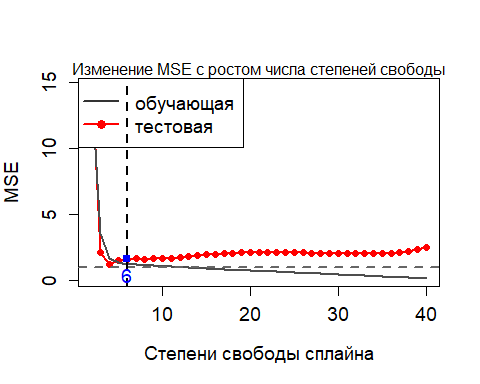
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 13.693854 | 14.731006 |
| 3 | 3.592171 | 2.157929 |
| 4 | 1.706881 | 1.226380 |
| 5 | 1.343484 | 1.526439 |
| 6 | 1.240088 | 1.636524 |
| 7 | 1.188622 | 1.641557 |
| 8 | 1.149483 | 1.632344 |
| 9 | 1.111930 | 1.637212 |
| 10 | 1.073470 | 1.661982 |
| 11 | 1.033861 | 1.705294 |

Расчёт ошибки на обучающей выборке:

Расчёт ошибки на тестовой выборке:

Изобразим на графике поведение ошибок при различном количестве степеней свободы.

###### График 2: Зависимость MSE от гибкости модели



На этом графике:

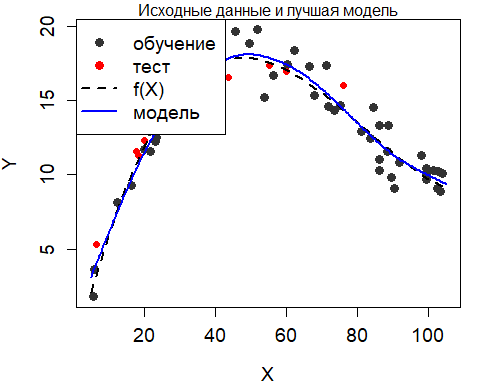
При движении слева направо MSE на обучающей выборке (серая кривая) очень плавно сокращается, потому что с ростом числа степеней свободы расчёт число узлов, по которым строится сплайн. При этом модельная кривая подгоняется по всё возрастающему количеству точек и становится всё более гибкой. В результате индивидуальные расстояния от фактических наблюдений за Y до их модельных оценок сокращаются, что приводит к сокращению MSE.

При движении слева направо MSE на тестовой выборке (красная кривая) сначала резко сокращается, затем очень плавно растёт. Нам известна истинная форма связи Y с X, она описывается кубической функцией. Число степеней свободы такой модели равно числу оцениваемых параметров. Поэтому резкое падение ошибки на тестовой выборке при небольшом числе степеней свободы связано с тем, что модель приближается по гибкости к истинной функции связи. Затем MSE на тестовой выборке довольно долго остаётся стабильной.

Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 3 и равно 1.16. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 6. Ошибка здесь равна 1, что ненамного отличается от минимума. Именно df = 6 было выбрано в качестве компромисса между точностью (минимальной MSE на тестовой выборке) и простотой модели (чем меньше степеней свободы, тем модель проще).

График с моделью, выбранной в качестве лучшей, показан на рисунке ниже.

###### График 3: Лучшая модель (компромисс между гибкостью и точностью)

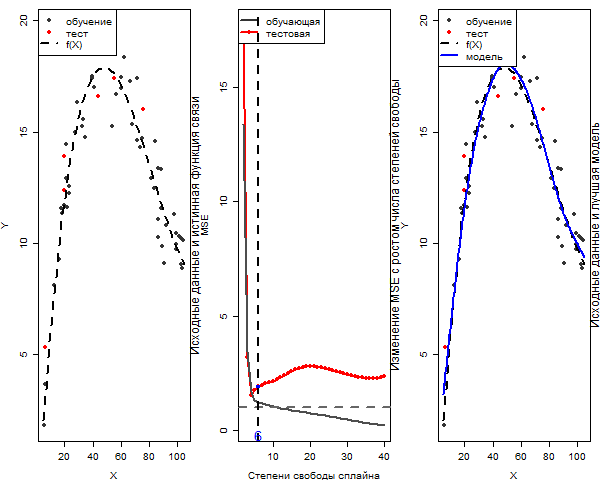


# Задача 2.

*Задание: Решить задачу 1, изменив характеристики данных. Почему при таком изменении данных MSE меняется именно так? Все рисунки сохранить в графические файлы в формате png.*

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 13.3835785 | 17.730047 |
| 3 | 3.4630081 | 3.167008 |
| 4 | 1.6701630 | 1.529200 |
| 5 | 1.3116033 | 1.740333 |
| 6 | 1.2013820 | 1.897148 |
| 7 | 1.1450377 | 1.968404 |
| 8 | 1.1041339 | 2.020538 |
| 9 | 1.0674511 | 2.077129 |
| 10 | 1.0311830 | 2.145029 |
| 11 | 0.9945274 | 2.225105 |

## png   
## 2

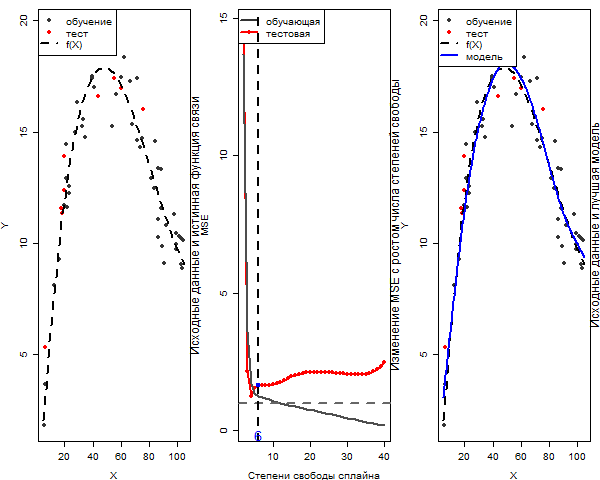


Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 10 и равно 1.860475. По графику мы можем установить, что первое значение MSEТЕСТ df = 8.

df = 8 было выбрано в качестве компромисса между точностью (минимальной MSE на тестовой выборке) и простотой модели (чем меньше степеней свободы, тем модель проще).

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 13.693854 | 14.731006 |
| 3 | 3.592171 | 2.157929 |
| 4 | 1.706881 | 1.226380 |
| 5 | 1.343484 | 1.526439 |
| 6 | 1.240088 | 1.636524 |
| 7 | 1.188622 | 1.641557 |
| 8 | 1.149483 | 1.632344 |
| 9 | 1.111930 | 1.637212 |
| 10 | 1.073470 | 1.661982 |
| 11 | 1.033861 | 1.705294 |

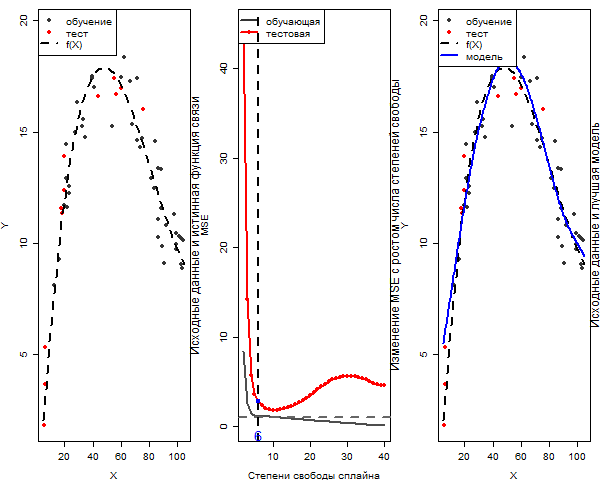
## png   
## 2



Наименьшее значение MSE на тестовой выборке df = 8 и равно 1.632344. По графику можно заметить, что первое значение MSEТЕСТ df = 6.

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 8.419357 | 44.780435 |
| 3 | 2.584743 | 14.213059 |
| 4 | 1.481283 | 5.737403 |
| 5 | 1.272348 | 3.662537 |
| 6 | 1.216545 | 2.863409 |
| 7 | 1.182827 | 2.390442 |
| 8 | 1.150180 | 2.096128 |
| 9 | 1.114940 | 1.930902 |
| 10 | 1.076638 | 1.860475 |
| 11 | 1.036721 | 1.861156 |

## png   
## 2



Наименьшее значение MSE на тестовой выборке df = 10 и равно 1.860475. По графику можно предположить, что значение MSEТЕСТ, соответствует df = 9.

* Почему при таком изменении данных MSE меняется именно так?

Изначально имеется доволно большой процент выборки и при ее уменьшении (доля обучающей выборки меняется от 0.9 до 0.8) так же имеет значение что модель пытается принять во внимание все имеющиесяя точки.